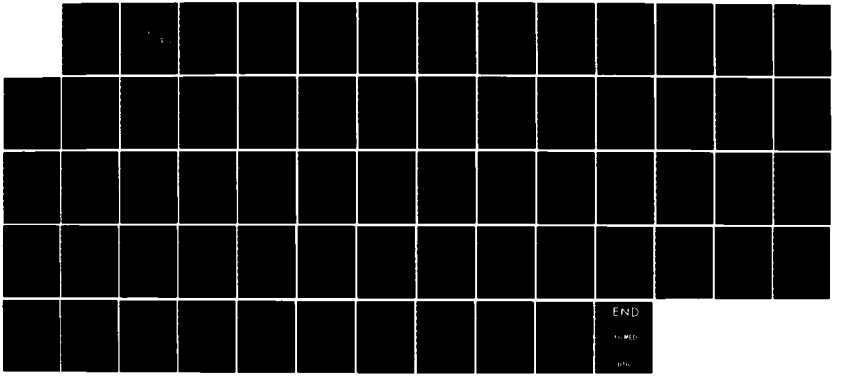
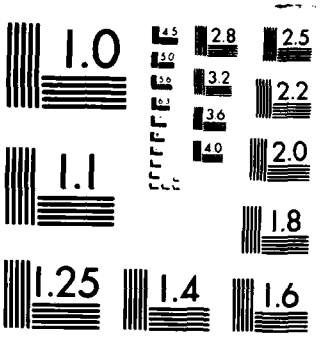
AD-A152 136 IMPLEMENTATION OF KOREAN AND CHINESE CHARACTERS THROUGH 1/1
COMPUTER(U) NAVAL POSTGRADUATE SCHOOL MONTEREY CA
C H KIM ET AL. SEP 84

UNCLASSIFIED F/G 9/2 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

IMPLEMENTATION OF KOREAN AND CHINESE CHARACTERS THROUGH COMPUTER

by

Chong Hae Kim
and
Sung Woo Ko

September 1984

Thesis Advisor: Michael J. Zyda

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A152 136	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* Implementation of Korean and Chinese Characters through Computer		5. TYPE OF REPORT & PERIOD COVERED Master's Thesis September 1984
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Chong Hae Kim and Sung Woo Ko		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, California 93943		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Postgraduate School Monterey, California 93943		12. REPORT DATE September 1984
		13. NUMBER OF PAGES 62
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

bit, byte, code, keyboard, keytop, bit-map, matrix printer, CRT, resolution, quality, enumeration

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Methods of representing Korean and Chinese characters are presented, using a limited number of keystrokes on a standard keyboard. Various attempts have been made to find the most efficient way to represent these characters such as enumeration methods, 16-bit coding for Korean character syllables, and the meaning and the sound method for Chinese characters. Details of these are explained with a brief introduction (Continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S N 0102-LF-014-6601

ABSTRACT (Continued)

to some general properties of Korean and Chinese characters currently used in Korea.

Approved for public release; distribution unlimited.

Implementation of Korean and Chinese Characters through Computer

by

Chong Hae Kim
Major, Republic of Korea Army
B.S., Korea Military Academy, 1976

and

Sung Woo Ko
Major, Republic of Korea Army
B.A., Korea Military Academy, 1976

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN INFORMATIONS SYSTEMS

from the

NAVAL POSTGRADUATE SCHOOL
September 1984

Authors: _____
Chong Hae Kim

_____
Sung Woo Ko

Approved by: _____
Michael J. Zyda, Thesis Advisor

_____
Carl R. Jones, Second Reader

_____
Willis H. Greer, Jr., Chairman,
Department of Administrative Sciences

_____
Kneale T. Marshall,
Dean of Information and Policy Sciences

ABSTRACT

Methods of representing Korean and Chinese characters are presented, using a limited number of keystrokes on a standard keyboard. Various attempts have been made to find the most efficient way to represent these characters such as enumeration methods, 16-bit coding for Korean character syllables, and the meaning and the sound method for Chinese characters. Details of these are explained with a brief introduction to some general properties of Korean and Chinese characters currently used in Korea.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

The development of computer and information processing has come to the stage of being able to handle Korean and Chinese character input and output. There is no problem in information systems for the input and output of characters from a standard Roman character keyboard, but the problems related to non-Roman characters from I/O to software problems of language handling remain almost unsolved. Until recently the computer could not handle Korean or Chinese characters efficiently. It was not user friendly and data processing in Korea was imperfect and very unwieldy. Among the problems, the biggest issue is how to enter 2,369 Korean and 1,800 common Chinese characters from the standard Roman character keyboard.

During the last few years, there have been great efforts at universities, research institutes and manufacturers for the development of good I/O devices for Korean characters. In Korea, natural language processing, especially Korean language processing, is one of the essential elements for the future of computer and information systems.

First the properties of Korean and Chinese characters will be presented as an introduction for those unfamiliar with these characters. Then, the resolution power of CRT's and dot matrix printers and their relation to the shape characteristics (readability, asthetic quality, etc.) of Korean and Chinese characters will be discussed. The methods which are developed for Korean and Chinese character I/O can be applied to other character sets, especially to many non-Roman alphabetic character sets, not to mention Chinese characters in China.

II. BACKGROUND

A. PROPERTIES OF KOREAN DOCUMENTS

Common documents in Korea are usually written in a mixed form utilizing Korean and Chinese characters. Minor use is made of Roman script. The usage of each character set depends on the kind of document. In order to perform word processing efficiently in Korea, the simultaneous editing of these characters is essential. Table I shows the use of characters found for various types of documents. This data is based on sampling performed expressly for this study. The following sources were in the sampling process to construct Table I:

1. newspaper - Korean Daily Times, "3A Era", 16 September 1984
2. journal - "National Security", June 1984
3. technical papers (A) - "COBOL Programming", Dong-A publishing Co., 1978
4. technical papers (B) - "Introduction to Law", Beob Moon Sa publishing Co., 1978
5. business papers - Korean Air Lines Co.

Although the sample was taken from a single source for each kind of document, it is the authors' view that the documents selected are representative of the entire population of each type.

B. CHARACTERISTICS OF KOREAN SCRIPT

The native Korean alphabet was introduced in 1446, after centuries of the use of a more cumbersome method (known as IDU) to transcribe Korean with Chinese characters. The set

TABLE I

Proportions of Written Characters

	Newspaper	Journal	Technical paper (A)	Technical paper (B)	Business paper
Roman script	1%	3%	40%	0%	10%
Korean script	84%	76%	55%	55%	80%
Chinese character	15%	21%	5%	45%	10%

* (A): Technical papers from western countries
* (B): Traditional and historical papers

of 28 letters[1] (now 24 letters) was designed by a group of scholars commissioned by King Sejong (1419 - 1450), the fourth King of the Yi dynasty.

The Korean language and alphabet is spoken and written by an estimated 50 million people on the Korean peninsula and its coastal islands. Many among the approximately one million Koreans residing in Japan, China, and America still speak and write the language [Ref. 9].

The Korean alphabet currently used consists of 14 consonants (ㄱ ㄴ ㄷ ㄹ ㅁ ㅂ ㅅ ㅇ ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ) and 10 vowels (ㅏ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅣ). There are also 17 compound consonants (ㄲ ㄳ ㄵ ㄶ ㄸ ㄺ ㄻ ㄼ ㄽ ㄾ ㄿ ㅀ ㅃ ㅄ ㅆ ㅉ) and 11 compound vowels (ㅐ ㅒ ㅔ ㅖ ㅘ ㅙ ㅚ ㅝ ㅞ ㅟ ㅢ). The letters of the Korean alphabet cannot be used independently but are used to build syllables. Each Korean character consists of two or three parts. The first part

[1] A letter is an element of a character. The character consists of two or three letters. Letters in Korea are a set of 14 consonants and 10 vowels.

must be a consonant or compound consonant. There are 19 letters that are possible for the first part of the Korean character. They are typically consonants or compound consonants (ㄱ ㄲ ㄴ ㄷ ㄸ ㄹ ㅁ ㅂ ㅃ ㅅ ㅆ ㅇ ㅈ ㅉ ㅊ ㅋ ㅌ ㅍ ㅎ). The second part of the Korea character is typically a vowel or compound vowel. There are 21 possible letters for the second part of the Korean character (ㅏ ㅐ ㅑ ㅒ ㅓ ㅔ ㅕ ㅖ ㅗ ㅘ ㅙ ㅚ ㅛ ㅜ ㅝ ㅞ ㅟ ㅠ ㅡ ㅢ ㅣ). The third part of the Korean character is optional and depends on the character being depicted. The third part if present, must be a consonant or a compound consonant. There are 28 letters possible as the third part (ㄱ ㄲ ㄳ ㄴ ㄵ ㄶ ㄷ ㄹ ㄺ ㄻ ㄼ ㄽ ㄾ ㄿ ㅀ ㅁ ㅂ ㅄ ㅅ ㅆ ㅇ ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ). This section has been summarized in Figure 2.1.

The Korean system of writing is called "Hangul". It is "phonetic" writing, like English, in the sense that the symbols represent sounds, that is, consonants and vowels. Unlike English symbols, which are grouped directly into words (e.g., E+n+g+l+i+s+h = English), Korean symbols are first grouped by syllable (e.g., H+a+n g+u+l = Han gul) [Ref. 10].

Korean symbols are written in syllabic groupings. An enumeration method[2] is to put letters side by side as in "LONDON". But the Korean language stacks the letters in most characters. For example, "LONDON" would be written 론 돈. The simplest syllable is written with one consonant and one vowel. When one writes the symbol for a vowel alone, one must add the consonant symbol "ㅇ", which indicates an initial mute (which is closed as a consonant). In this simple consonant and vowel syllable, there are two types of arrangements; side-by-side arrangement (e.g., 가) and

---

[2]In an enumeration method letters are placed side by side or element by element using a set of consonants and vowels.

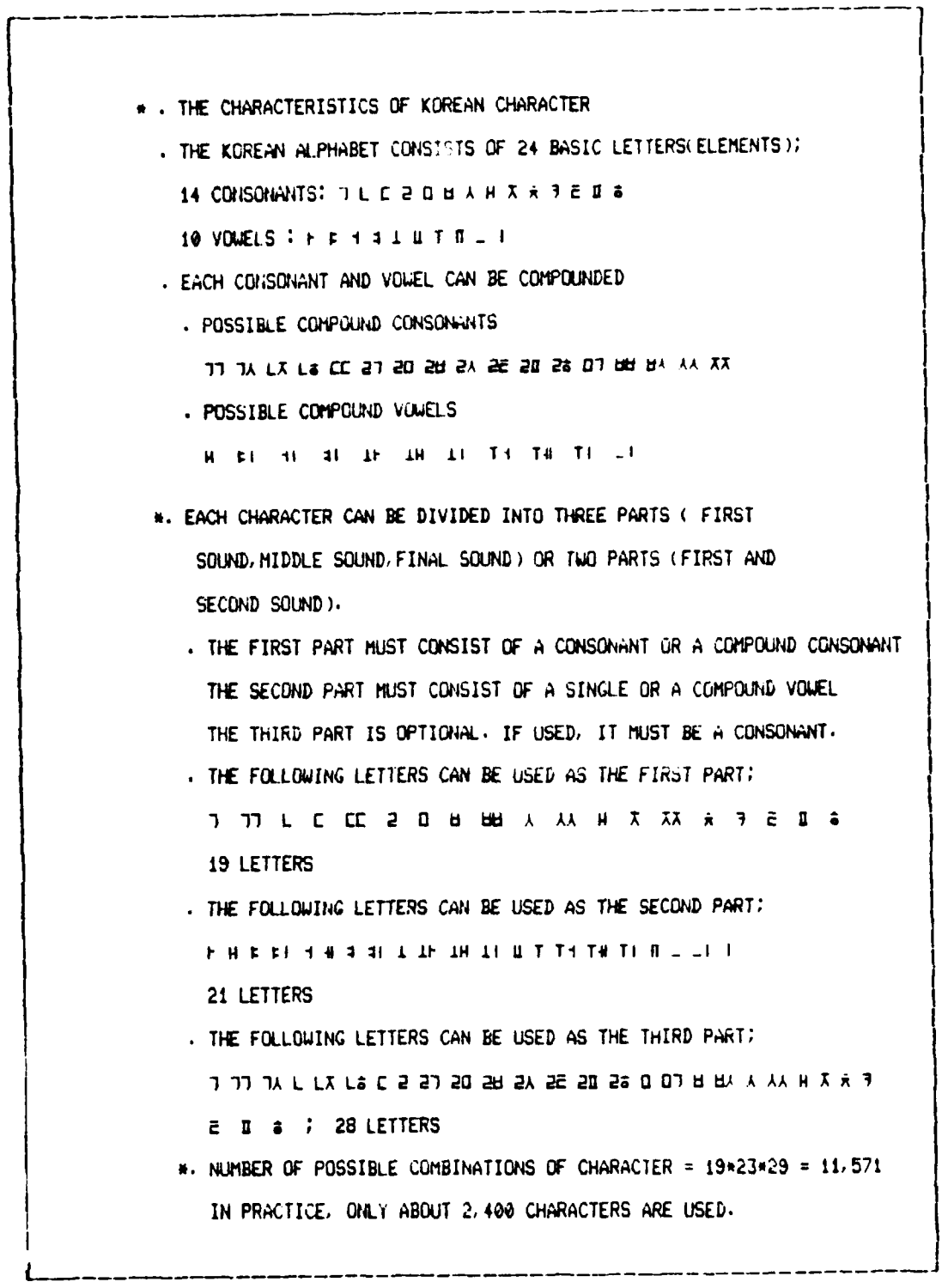


Figure 2.1 The Korean Alphabet.

top-to-bottom arrangement (e.g., 그). The particular vowel being written determines which arrangement is used.

Representing these character syllables through a computer creates a problem because each letter's (consonant and vowel) shape can be different due to a requirement that each character be balanced, i.e., have the same size and achieve a desired asthetic quality. For example, when ㄱ is placed to the left of a vowel, the downward portion is slanted: ㄱ (e.g., 가). When it is placed on top of the vowel, the downward portion becomes straight: ㄱ (e.g., 그). As shown above, it is very difficult to apply these different shapes for a particular letter to a line printer and a typewriter. This problem will be discussed in detail in the following chapter.

By mathematical calculation, the possible number of Korean characters is 11,571 (19 * 21 * 29). It must be noted through that only 2,369 characters are commonly used [Ref. 8: p. 11].

C. CHARACTERISTICS OF SINO-KOREAN CHARACTERS

Sino-Korean characters are Chinese characters used in Korea. They are different from those used in China. Koreans refer to Chinese characters as Hanja. Chinese characters have a long history, the earliest discovered writings having been dated from about 14 B.C.. In 100 A.D. during the Han Dynasty, this was modified by Hsu Sheng (許愼, 30 - 124) in his 15 - Volume paleographical work, Shuo-wen Chieh-tzu, (說文解字) which translates to the explanation of writing and analysis of words. That work lists 9,353 characters under 540 radical entries. Of this number, 364 are pictographic, 125 simple idiographic, 1,167 compound idiographic and 7,697 phonetic ccmpounds.

The most complete collection, the Kang Hsi Dictionary with about 50,000 characters was published in 1716. Since 1949, after the establishment of the Peoples Republic of China, the Chinese government actively pursued language reform until the Cultural Revolution, 1966-1976. The Chinese government changed and simplified the characters from the original [Ref. 5: p. 15].

The number of characters used commonly is from 1,000 to 3,000. Table II [Ref. 1: p. 819] shows the frequency of

TABLE II

Frequency of Chinese Characters Used in Documents

	Newspapers (%)	General Document (%)	Total Document (%)	Newspapers (chrs)	General Document (chrs)
1st 10 chrs	10.0	8.8	80	499	638
50	27.5	25.5	85	615	777
100	38.9	36.1	90	781	992
200	55.4	51.0	95	1068	1358
500	79.0	73.5	96	1156	1479
1000	93.1	89.0	97	1269	1617
1500	97.4	95.0	98	1421	1832
2000	98.7	97.6	99	1661	2157
2500	98.9	99.4	100	2879	3328
3000		99.8			

* chrs: acronym of characters

Chinese characters used in typical documents.

In 1972, the Korean ministry of Education suggested that 1,800 Chinese characters be learned and used for educational purposes [Ref. 3]. In this study, the authors will restrict themselves to that set of 1,800 characters. The Chinese characters are called Hantzu in Chinese, Hanja in Korean, and Kanji in Japanese. All mean "Han Characters" (漢字).

These characters are used exclusively in Chinese writings, and in combination with the Hangul (Korean) alphabet in Korea and with the Kana Syllabaries in Japan. The Sino-Korean (Hanja), in written form, is a combination of three major elements: pictograms and ideograms, and phonograms [Ref. 5: p. 22].

In the next chapter the perspective of a picture for each character will be used because of both the complexity of Chinese chracters and the ease of representation in the computer. Each Chinese character has the meaning and sound, for example, 天 means heaven and the sound is cheon. Also, there are many characters which have different meanings but the same sound, or the same meaning but different sounds. In order to solve this problem there are several methods. Appendix A [Ref. 5: p. 17] shows the evolution of Chinese characters.

III. PROBLEMS OF EDITING KOREAN AND CHINESE SCRIPTS

A. CURRENT EDITING TECHNOLOGY

The current word processing practice in Korea is to type Korean characters by the enumeration method, that is, input letters (8 bit code: consonant and vowel in sequence) and output these letters as a character syllable using a Korean character conversion program for Korean script. Appendix B shows the EBCDIC input codes currently used by FACOM, and Appendix C depicts MDS (Mahawk Data Sciences) input codes used by IBM. To type Chinese characters the following sequence is followed:

1. Depressing a Chinese character function key.
2. Typing the sound character of a Chinese character using the enumeration method.
3. Displaying all homonym (from 1 to 60) characters [Ref. 4: p. 34] that have the same sound.
4. Selecting one character by using an index number, and entering the character to a buffer or file.

Machines dealing with Korean language data are currently available from the IBM and FACOM corporations in Korea; IBM's Multistation 5550 (1984) and FACOM OS IV(KEF) (1982) are newly updated and well developed machines. These machines still have several disadvantages in handling Korean and Chinese characters:

1. A large amount of time is spent in character conversion.
2. It is difficult to directly delete and insert records in a file.
3. The word processing editor cannot recognize the characters being edited before executing a character

conversion program since only the enumerated letters can be displayed.

4. The method of entering characters is inconvenient and requires a tremendous amount of effort for Chinese characters.
5. One cannot convert all Korean character syllables into Chinese characters because there is not a one to one mapping.
6. Data communication is impossible since there are no standard codes for Korean and Chinese characters.

Appendices D and E show the keyboard of IBM Multistation 5550 [Ref. 8: p. 14] and FACOM OS IV(KEF) [Ref. 7: p. 48] respectively.

B. USER REQUIREMENTS

Most potential users have recognized that the computer is essential in data processing and office automation. However, because of the above constraints, they are unsatisfactory for use with the Korean language. Some general user requirements of computer researchers and manufacturers are the follcwing:

1. Users want to use Korean language commands and programs but there are no Korean language oriented operating systems or programming languages such as COBOL, FORTRAN, Pascal, etc.
2. Users want to edit three kinds of characters simultaneously and in a user friendly manner.
3. Users want to display and print out data without using a conversion program, as is done with the Korean alphabet because of time, memory space, and inconvenience.
4. Users want to use interactive files and database processing.

In summation, they want to use computers that handle three kinds of script in the same manner in which present computers do with the Roman alphabet.

C. PROBLEMS OF REPRESENTATION OF THE THREE KINDS OF SCRIPTS

Because of the characteristics of Korean and Chinese characters, the following problems occur:

1. How can one enter 2,400 Korean characters and 1,800 Chinese characters into a computer through a limited number of keystrokes.
2. How can one develop the system program to direct input and output without using a conversion program.
3. How can the asthetic quality of display and output be improved.
4. How can one increase the processing speed and reduce the memory space for these character definitions.

There are other problems but the above problems are the most significant. Among these problems the first one is the most serious and significant problem, and consequently, the authors will give it more attention in this study.

IV. POSSIBLE METHODS FOR KOREAN LANGUAGE DATA PROCESSING

In order to solve the problems which were mentioned in the previous chapter, the following methods are offered as possible alternatives for Korean language data processing.

A. 8-BIT CODE FOR KOREAN ALPHABET

Since the Korean alphabet consists of only 24 letters and Korean language data can be expressed using only Korean characters without a serious problem. The enumeration method, like the Roman alphabet, is the easiest way to represent Korean characters without changing the hardware and the operating system. This method is not highly readable and would require changes in the language which may not be acceptable to users.

1. Using the Current Standard Keyboard

A program can be loaded which defines the 24 letter Korean alphabet to a character generator instead of the lower case Roman alphabet. All Korean alphabet elements and the upper case Roman alphabet characters are then available through the standard Roman character keyboard. With this method the user can use a computer in a similar manner as the users who use the Roman alphabet. In addition, well developed hardware and software can be used without critical problems. This method has been suggested by many groups of people from the time when the Korean typewriter was first developed. The only disadvantage is the breaking of traditional custom. To capitalize on developed technology and for the ease of application, more study and research should center on user acceptability of the enumeration method.

Figure 4.1 shows an example of hard copy which uses a graphic dot printer and a standard keyboard. Appendix F shows the load command program for an alternative character

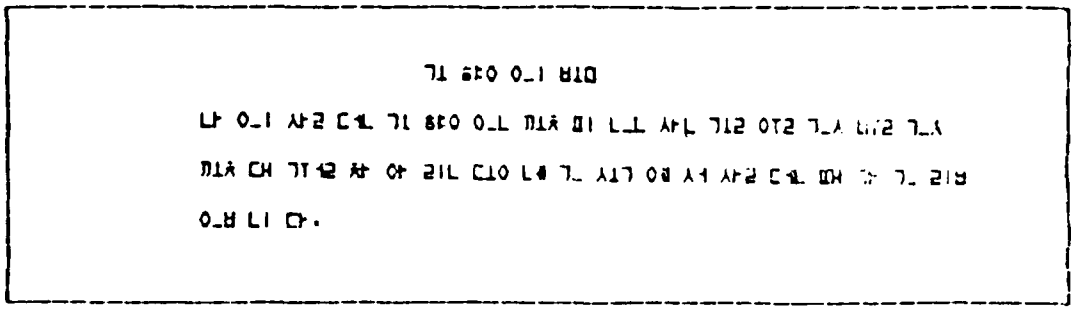


Figure 4.1 Example Using Standard Keyboard.

generator for the Korean alphabet. This program can be generated easily by the alternative character set editor, and it loads the Korean alphabet to an alternative character generator instead of the lower case of Roman alphabet.

2. Using the Capital Letters as the Initial Letter

The major difficulty with the enumeration method is poor readability. Korean users read a sentence sequentially syllable by syllable. In order to increase readability, the initial letter of each character can be written as an upper case letter to distinguish the syllable easily. Figure 4.2 shows the example using the capital letters and Appendix G represents the load command program for these letters. A special mark or altered shape of each letter also can be applied to increase a readability when an enumeration method is used.

Figure 4.2 Example Using Capital Letters.

B. 16-BIT CODE FOR THE THREE KINDS OF SCRIPT

There are various methods one can use to enter Korean and Chinese characters, but the 16-bit code is one of the letter methods, since it can identify all possible Korean and Chinese characters without using the enumeration method and a conversion program. The structure of this code will be discussed briefly in the following subsection.

1. 16-bit Code for Korean Script

As mentioned before, a Korean character syllable consists of three parts:

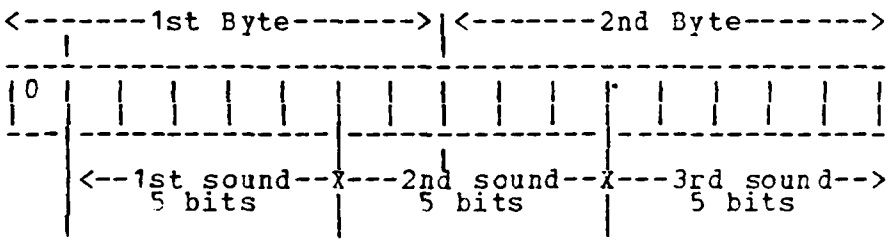
1. First sound; one of 19 simple or double consonants.
2. Second sound; one of 21 simple or compound vowels.
3. Third sound; one of 28 simple or compound consonants (optional).

Since the number of each first, second, and third letters is less than 32 letters, 5 bits are enough to identify each sound. All possible Korean characters can be identified using 15 bits. The 1st bit of 16 bits is used to indicate a Korean character (by a 0). The next 5 bits are used for the first sound, the following 5 bits for the second sound, and the final 5 bits for the third sound. Table III shows the structure of 16-bit code for the Korean

character and Table IV explains the 16-bit code for Korean character. This code table is basically the same as the IBM 2-byte internal Korean character code [Ref. 6: p. 52]. The only difference is the arrangement. Some IBM codes represent three letters. This makes key tops (face of each key) more complex; for example, 00100 (one key top) represents ㄴ, ㅐ, and ㄱㅅ values. The code suggested in Table IV reduces some of this complexity by limiting the possible values to no more than two for each keytop. In contrast to the example for IBM codes, the same code from Table IV represents only one value. Appendix H represents the IBM 2-byte internal

TABLE III

Structure of 16-bit Code for Korean Script



* 0: Korean character

Hangul code for the Korean character.

The suggested code has several advantages. First, it is easy to sort the character order by its value since the value of each letter is in the order of the Korean alphabet. Second, it can reduce the memory space for data by using 2 bytes instead of 3 bytes for one character. Third, it is possible to edit the character directly since it does not need code conversion. Finally, since it can easily

recognize the code value of the Korean character, it helps a programmer when it is programmed.

TABLE IV

16-bit Code for Korean Script

5 bit code	1st sound letter	2nd sound letter	3rd sound letter	5 bit code	1st sound letter	2nd sound letter	3rd sound letter
00000				10000		ㅐ	ㄶ
00001	ㄱ		ㄱ	10001	ㅁ		ㅁ
00010	ㄲ	ㅏ	ㄲ	10010		ㅛ	ㅁㄱ
00011		ㅐ	ㄳ	10011	ㅂ		ㅂ
00100	ㄴ		ㄴ	10100	ㅃ	ㅜ	
00101		ㅑ	ㄵ	10101		ㅝ	ㅄ
00110		ㅒ	ㄶ	10110	ㅅ		ㅅ
00111	ㄷ		ㄷ	10111	ㅆ	ㅞ	ㅆ
01000	ㄸ	ㅓ		11000	ㅇ		ㅇ
01001	ㄹ		ㄹ	11001	ㅈ		ㅈ
01010		ㅔ	ㄺ	11010	ㅉ	ㅟ	
01011		ㅕ	ㄻ	11011	ㅊ	ㅠ	ㅊ
01100		ㅖ	ㄼ	11100	ㅋ	ㅡ	ㅋ
01101		ㅗ	ㄽ	11101	ㅌ	ㅢ	ㅌ
01110		ㅘ	ㄾ	11110	ㅍ	ㅣ	ㅍ
01111		ㅙ	ㄿ	11111	ㅎ		ㅎ

* Blank: Not used

2. 16-bit Code for Chinese Characters

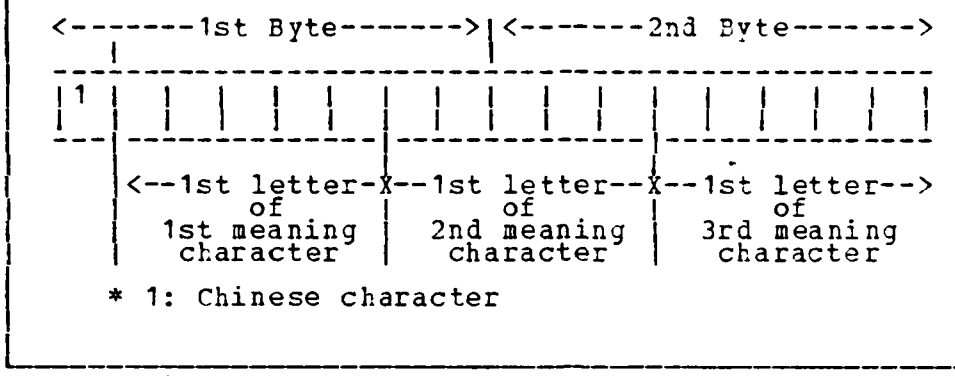
There is no limitation in the number of usable Chinese characters, but statistics show that 1,800-3,000 characters cover 98-99.8 percent of those which appear in newspapers and journals (Table II). Currently there are only two ways to represent Chinese characters in Korea. One method is comprised of two steps. The first step is to display all Chinese characters (synonym) which have the same sound after entering the desired sound, and the second step is to enter the Chinese character which is needed by the user via an index number matched to that character after selecting it in the display. The other method is to convert a Korean character to a Chinese character after typing a Korean character as a unit of a word, which consists of two or three characters.

The former is inconvenient and takes a long time to edit. The latter has no flexibility in that it is limited by the programmed word codes. To solve the above problem and simplify the identification of each character using a limited number of keystrokes, a 16-bit code for Chinese characters can be applied. Table V represents the structure of 16-bit code for Chinese characters.

Chinese characters represent both meaning and phonetics to Koreans. To simplify the code, all the complete meaning and sound of the Chinese characters are not needed. The Chinese characters are composed of from one to five syllables for meaning and one character for the sound. Simplicity can be achieved by employing abbreviations or acronyms for each part (meaning and sound). For example, a Chinese character (天) has a meaning as "Hea-Ven" and sound as cheon. In this case we use H of Hea, V of Ven, and C of cheon as a code for (天).

TABLE V

Structure of 16-bit Code for Chinese Characters



* 1: Chinese character

But this method may result in duplicate codes for different Chinese characters which mean another character and may have the same value as HVC. In order to eliminate the duplicate code and to use the 3 letter code which is compatible with the 16-bit Korean character code, the following characteristics of the sound and meaning of Chinese characters are relevant: First, only 428 syllables are used to represent the sounds for all Chinese characters. That is, one sound can represent 1 to 60 Chinese characters. Second, the frequency of Korean characters used for the meaning and sound is irregular in distribution. More specific, 20% of Korean characters are used to represent the sound and meaning of 95% of Chinese characters [Ref. 11].

As a result of analyzing the 1,800 sound characters and 1,438 meaning characters used to represent the Chinese characters, Table VI and Table VIII are derived. Table VI represents the number of Chinese characters which have the same first sound letter and the same second sound letter. For example, 266 Chinese characters have the first sound letter (_ㄱ_), 44 Chinese characters have (_ㄱ_) as a first

TABLE VI

Characters Having Same 1st & 2nd Letter Sound

2nd letter		1st letter												
		ㄱ	ㄴ	ㄷ	ㄹ	ㅁ	ㅂ	ㅅ	ㅇ	ㅈ	ㅊ	ㅌ	ㅍ	ㅎ
ㅏ Group	ㅏ	44	8	26	17	20	27	62	25	41	22	17	11	27
	ㅐ	12	6	8	4	13	12	4	5	11	7	8	2	9
	ㅑ				9				17					5
	ㅒ													
ㅓ Group	ㅓ	18		1			11	42	12	59	28			7
	ㅔ	1						6		14	2			
	ㅕ	34	4		26	12	12		43				7	18
	ㅖ	15							4				6	3
ㅗ Group	ㅗ	34	5	36	13	18	21	28	16	28	15	9	10	24
	ㅘ	22							7	4				24
	ㅙ	1						2						
	ㅚ	4	2		2			1	2	1	2	1		8
	ㅛ	8			3	5			13				4	3
ㅜ Group	ㅜ	32		6	5	16	33	44	22	22	18	5	4	6
	ㅝ	6							13					
	ㅞ													1
	ㅟ	3							14		6			2
	ㅠ	5			12				23					4
ㅡ Group	ㅡ	20	1	4	1			10	12	8	4	1		3
	ㅢ								10					6
	ㅣ	16	1		13	14	19	37	29	36	18		9	
Subtotal		263	27	81	105	98	135	237	265	224	122	41	53	148

Total 1,800

sound letter and (___) as a 2nd sound letter. One must rearrange the sound acronym to the 5 bit code since the distribution of sound characters is irregular. Table VII depicts the rearranged code value for acronym of the Chinese sound character.

TABLE VII

5-bit Code for the Acronym of Sound Character

code	sound acronym	code	sound acronym
00000	ㄱ	10000	ㅅ
00001	ㄲ	10001	ㅏ
00010	ㄱ	10010	ㅓ
00011	ㄱ	10011	ㅗ
00100	ㄱ	10100	ㅜ
00101	ㄴ	10101	ㅡ
00110	ㄷ	10110	ㅈ
00111	ㄸ	10111	ㅓ
01000	ㄹ	11000	ㅈ
01001	ㅁ	11001	ㅈ
01010	ㅂ	11010	ㅈ
01011	ㅃ	11011	ㅊ
01100	ㅅ	11100	ㅌ
01101	ㅅ	11101	ㅍ
01110	ㅅ	11110	ㅎ
01111	ㅅ	11111	ㅎ

In Table VII the second letter (_ㅏ_,_ㅓ_,_ㅗ_,_ㅜ_,_ㅡ_) describes all the group letters. For example, (_ㅏ_)

represents _ㅏ_, _ㅐ_, _ㅑ_, _ㅒ_, and (_ㅓ_); _ㅔ_, _ㅖ_, _ㅕ_, _ㅖ_, and (_ㅗ_); _ㅘ_, _ㅙ_, _ㅚ_, _ㅛ_, etc. Also (_ㄹ·_) represents _ㅏ_, _ㅓ_ group which assembles the first consonants at the left of the vowel. (_ㄹ·_) describes _ㅗ_, _ㅜ_, and _ㅡ_ group which assembles the first consonants above the vowel.

Since the frequencies of Korean character syllables representing Chinese character's sound and meaning are different, the frequency of Korean characters to represent Chinese characters meaning is needed to be analyzed. After analyzing the sampled 1,438 characters which are the first and the second meaning characters, Table VIII is derived which shows the number of times for a meaning character or a group to be used.

The meaning acronym value to a 5-bit code from the basis of Table VIII can be reassigned. Table IX shows the reassigned 5-bit codes representing the acronym of the meaning character. The same theory can be applied as in Table VII when Table IX is derived. As the acronym code is rearranged, the proportion of the duplicate codes can be reduced. As a result of applying these rearranged codes, Table X can be produced which shows the proportion of the duplicate codes. The pure acronym code (Table IX) represents the acronym of a meaning and a sound character as a first letter code of Korean characters (_ㄱ ㄲ ㄴ ㄷ ㄸ ㄹ ㅁ ㅂ ㅃ ㅅ ㅆ ㅇ ㅈ ㅉ ㅊ ㅋ ㅌ ㅍ ㅎ_; 19 possible consonants), the arranged sound character acronym code (Table VII), and the arranged sound and meaning character acronym code (Table VIII).

The reasons why some duplicate codes cannot be eliminated are: First, some Chinese characters have similar meaning and sound which generates the same acronym code (22 among 1,800); and second, there are initially some Chinese characters which have the same meaning sound (12 among

TABLE VIII

Frequency of Meaning Character in Chinese Characters

1st lt	2nd letter													
	ㅏ group		ㅓ group			ㅗ group			ㅜ group			ㅡ group		TOT
	ㅏ	ㅐ	ㅓ	ㅔ	ㅖ	ㅗ	ㅘ	ㅚ	ㅜ	ㅝ	ㅟ	ㅡ	ㅣ	
ㄱ	42	6	25	9	22	25	6	6	27	3	9	26	30	236
ㄲ	1	4	2			2		1	4	1		4	4	23
ㄴ	23	8	7		3	9		1	7			13	2	73
ㄷ	23	9	12			24	2	1	17		1	11	3	103
ㄸ	1	1	3	2					2		2	1	1	13
ㄹ	10	6	5	3	7	10		2	4		1	28	31	107
ㅁ	39	13	10	1	3	17		2	42			1	4	93
ㅂ	24	5	11		13	11			15				10	89
ㅃ	2	2	1		2				1			1		9
ㅅ	32	11	21	8		15		4	17		1	11	19	107
ㅆ	5		1			15	1	2				4		10
ㅇ	28	6	29	4	19	15	1	2	22	4	2	57	55	246
ㅈ	18	5	8	2		6			12			3	39	93
ㅉ	2					1			1				4	8
ㅊ	5	4	6			3			4			5	13	40
ㅋ	2	1				1					1	9		14
ㅌ	8	3	4			3			2			2	1	23
ㅍ	4	2	1		1	1			5			1	2	17
ㅎ	32	2	2	2		6	7					2	10	69
	301	87	149	31	70	143	16	19	182	8	17	187	228	1438

* 2nd letters having 10 are omitted for simplicity

TABLE IX

5-bit Code for Acronym of Meaning Character

code	1st & 2nd meaning acronym	code	1st & 2nd meaning acronym
00000		10000	ㅅ·
00001	가	10001	숫
00010	거	10010	아
00011	고	10011	어
00100	구	10100	오
00101	그	10101	우
00110	ㄴ·	10110	으
00111	눈	10111	ㅈ·
01000	ㄷ·	11000	주
01001	둔	11001	차
01010	ㄹ·	11010	ㅋ
01011	른	11011	ㅌ
01100	ㅁ·	11100	ㅍ
01101	문	11101	ㅎ·
01110	ㅂ·	11110	흔
01111	분	11111	

1,800). For these characters which have duplicate codes the users must apply the exception rule. Alternatively, the meaning of the character can be redefined to a synonym with a different acronym. For example, if the wrong homonymous Chinese character is displayed, the input operator may select another form of the homonym by keying in the full sound syllables instead of the acronym.

TABLE X

Proportion of Duplicate Code

Number of character	pure acronym code	Rearranged sound character code	Rearranged sound and meaning character code
1,800	7.5%	2.3%	1.8%

3. 16-bit Code for Roman Alphabet and Symbols

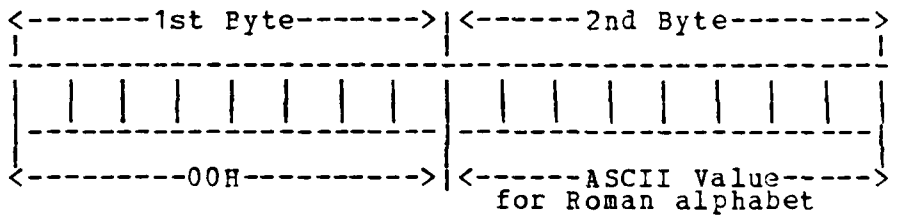
In order to use the three mixed kinds of a character code, and simplify the I/O controller, and unify the word, 16-bit codes for the Roman alphabet and symbols should be generated by only one keystroke. For data communication and for familiarity, adding only the default byte (00H) to ASCII code, 16-bit code for Roman alphabet, symbols, and control characters can be defined. When one uses only Roman alphanumeric characters, one can easily convert this 16-bit code to ASCII code. Table XI shows the 16-bit code for Roman alphabet and symbols.

4. Keyboard for 16-bit Code

As it is mentioned in the previous chapter, the biggest issue is how to enter all Chinese characters, Korean characters, and the Roman alphabets with a simple keyboard. In order to implement the 16-bit code to keyboard, one would have to make the keystrokes which generates "1" or "0" as a Chinese character function key (bit 1), three 5-bit codes (00000-11111) for Chinese and Korean characters (bits 2-16) and 16-bit code for Roman alphabet and symbols. In this case 33 more keys than the common Roman alphabet keyboard are

TABLE XI

Structure of 16-bit Code for Roman Alphabet

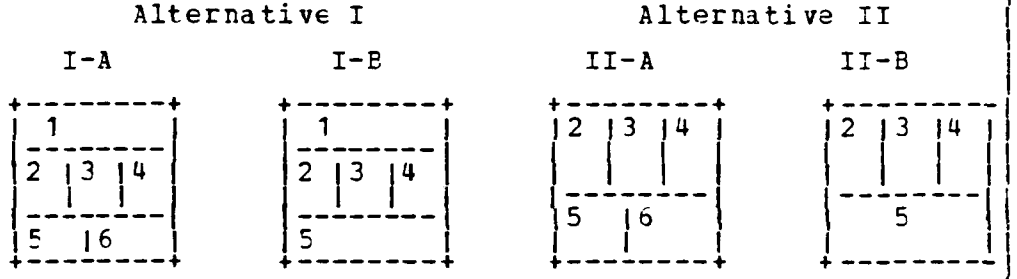


needed. To reduce the number of keys, one more function key can be added for the Roman alphabet which generates 16-bit code as a 5-bit code key. However, the user identification will be complex because there must be 4 or 5 letters on each key top. Table XII explains the 4 alternatives. Alternative I includes 32 Roman alphabet on 5-bit key tops using a Roman alphabet function key and Alternative II excludes Roman alphabet on 5-bit key top. Alternative A uses the acronym of sound and meaning characters and Alternative B uses only the acronym of sound character for the sound and meaning characters.

To select the best one, the authors can use the following criteria: flexibility of hardware and software design, hardware efficiency, ease of maintenance, system reliabilty, user characteristics, number of keystrokes, number of duplicate codes, and complexity of recognizing a certain keystroke [Ref. 2]. In the opinion of authors' alternative II-B should be selected for the ease of explanation and understanding. By selecting Alternative II-B, a user can type the three characters simultaneously. For example, to type "School is 학교 (Hak-kyo) in Korean and 學校 (Hak-kyo) in Chinese character", one can type directly

TABLE XII

Leveled Letters on 32 Key Tops



Legend:
1 : Roman alphabet
2 : First letter of Korean character
3 : Second letter of Korean character
4 : Third letter of Korean character
5 : Acronym of Chinese sound character
6 : Acronym of Chinese meaning character

only Roman alphabets in the previous sentence without a function key. Two syllables of "school" are formed from: In Korean, the first syllable is "ㅎ" selected from the position of first sound letter, "ㅏ" from the second sound position, "ㄱ" from the third position. The second syllable is "ㄱ, ㅛ, Default". Then, in Chinese, press the Chinese character function key which generates "1" as the first bit. The first syllable is "ㅂㅏ" which is the acronym of first meaning character and "ㅜ" which is the acronym of second meaning character and "ㅎ" which is the acronym of sound character. The second syllable is "ㅎㅏ ㄱㅛ ㄱㅛ". After typing Chinese characters, user must release the function key to type Korean characters. Table XIII explains the above example.

As a result of the above example, the computer generates the following codes in hexadecimal: 0053(S),

TABLE XIII

Typing Procedures for Mixed Characters

Tc type "schocl, 학 교, 學校", the followi ng procedures should be follwed:

1. Type "schocl" by one keystroke for each character without a function key.
2. In Korean, to type "학 교",
   first, type "ㅎ, ㅏ, ㄱ",
   second, tyre "ㄱ, ㅛ, Default".
3. In Chinese, to type "學 校", at first, press the functicn key, then type "ㅂ·, ㅜ, ㅎ·", "ㅎ·, ㅛ, ㅛ" (Table IX) since for "學", the meaing is "배 울" and the sound is "학", and for "校", the meaning is "학 교" and the sound is "교".

0063(c), 0068(h), 006F(o), 006F(o), 006C(l), and 0(Korean character), 11111(ㅎ), 00010(ㅏ), 00001(ㄱ); that is 0111 1100 0100 0001(7C 41:학), 0 (Korean character), 00001(ㄱ), 10010(ㅛ), 00000(Default); that is 0000 0110 0100 0000(06 40:교), and 1 (Chinese character), 01010(ㅂ·), 10100(ㅜ), 11110(ㅎ·); that is 1010 1010 1001 1110(AA 9E:學), and 1(Chinese character), 11110(ㅎ·), 00010(ㅛ), 11110(ㅛ); that is 1111 1000 0101 1110(F8 5E:校).

5. Operating System for Input and Output

To apply the suggested system, it is needed to redesign the operating system for input and output control.

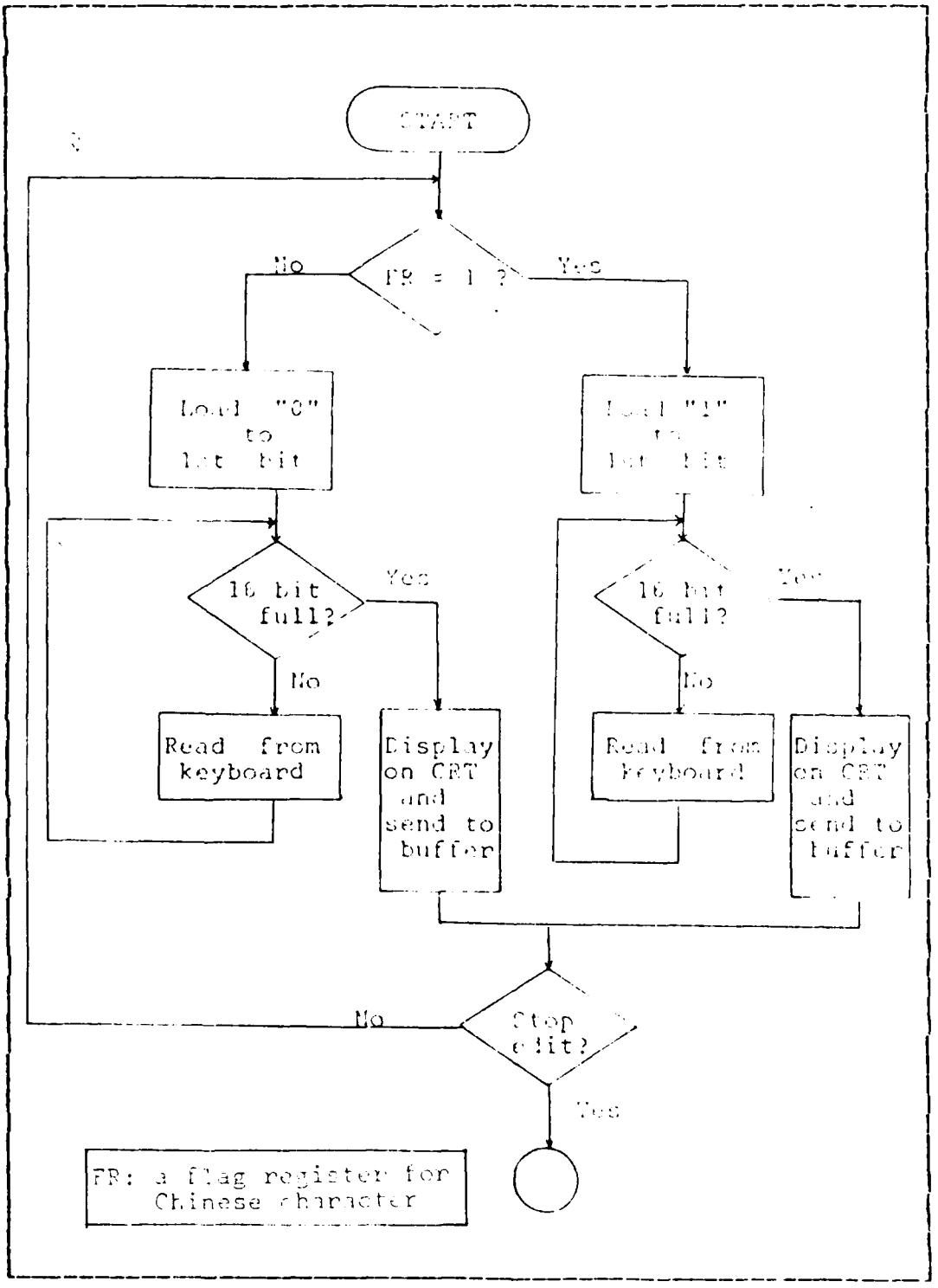


Figure 4.3 Flowchart of Input and Output Controller.

Figure 4.3 shows the flowchart of input and output control. First, the input and output controller has to distinguish whether the Chinese character function key is "0" or "1". A flag register can be used for Chinese character function key. For example, if the flag is "1", then "1" is loaded to the first of the 16 bit register, and multiple 5-bit codes are read until the register is full. If it is full, a character is displayed on the CRT, and the 16-bit code is sent to a buffer as data. Otherwise the flag is "0", and then "0" is loaded into the register, and 5-bit codes are read for a Korean character. A 16-bit code is used for a Roman alphabet character or a symbol code until the 16-bit register is full. If the 16-bit register is full, the identified character is displayed and sent to a buffer as data. If the "stop edit?" condition shown in Fig 4.3 is "no", the input and output controller makes a loop to read a code, displays a character and sends a character code to a buffer.

This system will make the use of Korean language commands and programs easier to use than those presently available. To achieve the above goals, a compiler and interpreter, as well as the operating system will require redesigning. This system will require the complete rewriting of all software currently used. The economic impact of this on the Korean people will be enormous.

6. Design Considerations for Character Generation

There are two shapes of characters used in Korea: Gothic (Figure 4.4) and Brush type (i.e., Ming style: Figure 4.5) [Ref. 7: p. 34]. To generate the above shapes of characters, several methods of a character generation can be considered. To select the best method for Korean and Chinese characters, one can use the following five criteria: speed, space, quality, flexibilty, and cost. Speed is a double standard: speed of creation may range from a few minutes to

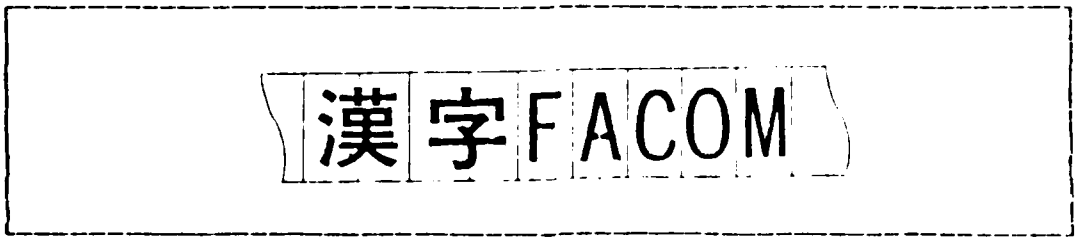


Figure 4.4 An Example of Gothic Type.

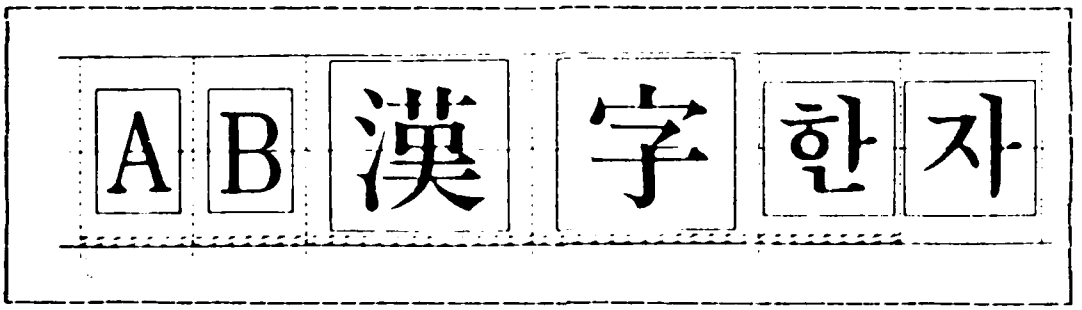


Figure 4.5 An Example of Brush Type.

a few hours, while speed of production should go beyond 1000 characters/sec depending on type size and device resolution. Space refers to the average size of the code for one character as well as the size of the internal buffers often needed for decoding. Quality is proportional to the largest dot matrix which can be used to decode a character; it should not be confused with the resolution of the output device. For a given type size, the resolution sets the definition, that is the size of the matrix to be used; definition, hence type size, is bounded by the quality of the code. Flexibilty refers to the different automatic modifications which are supported by the code: scaling, rotating, family variations (as going from light to bold). Cost is selfexplanatory [Ref. 12: p. 240].

Obviously the five criteria above are not independent. Figure 4.6 shows the interrelationship of criteria in designing a character generator [Ref. 12: p. 241]. The most desirable feature is indicated by the direction of the arrow. Solid (resp. dotted) lines indicate agreement (resp.

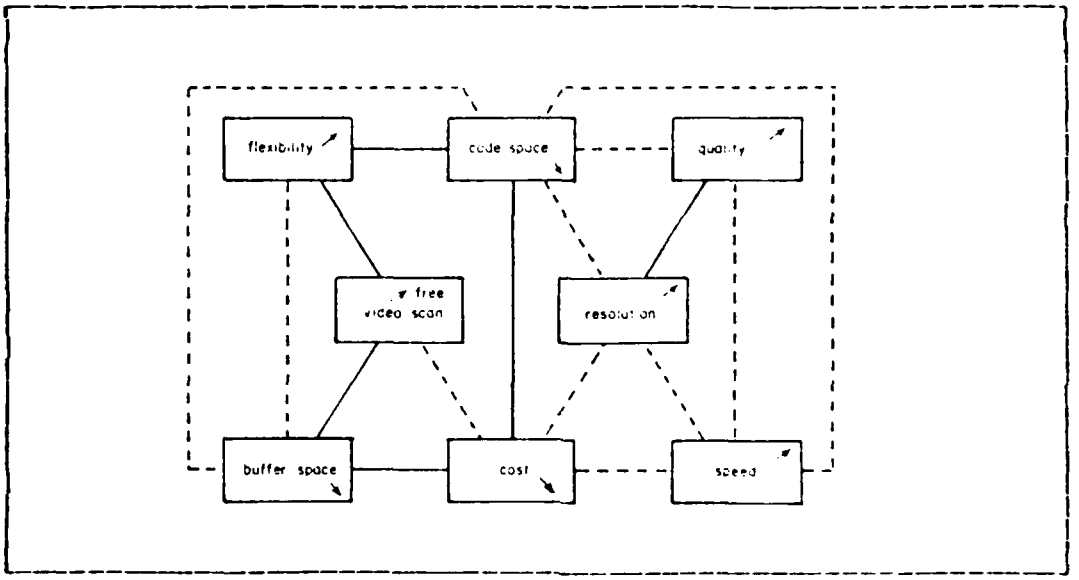


Figure 4.6 Criteria in Designing a Character Generator.

contrariety) between the variation of the factors. The design of a digital character generator is an engineer's task whose goal is to strike the appropriate balance between the specifications for those five criteria combined with the characteristics of the production device, resolution and scanning, and the necessity of operating the corresponding creation station.

Table XIV [Ref. 12: p. 268] gives a summary of the main characteristics of the coding methods that the engineer can utilize [Ref. 12: p. 269]. As the characteristics of Korean and Chinese characters are compared to Table XIV, it should be apparent that the bit map method is the most

TABLE XIV

Comparative Table for Performance

Method	Code space (bits)	Buffer space (bits)	Flexibility	Video scan	Resolution (n)	Qty	Spd
Bit map	n^2	0	---	+ +	≤50	--	+++
Run-length	$k*n*\log_2 n$	n	---	+ +	≤100	--	++
Chain-link	$6*k*n$	n^2	---	---	≤100	--	+
Differential run length	$6*k*n + b(\log_2 n+c)$	n	-	+	≤100	-	+
Spline	$k*\log_2 n$	$k_m*\log_2 n$					
Structural	$k'*\log_2 n$	n^2					

* Qty: Quality
* Spd: Speed

Legend:
+: Good
-: Bad
b: The numer of birth point in the character
c: A constant taking care of bookkeeping
n: The size of matrix
k: The average number of runs per matrix line or column (a number of the simplicity of the character shapes: approximately 4 for Roman body-text fonts, higher than 10 for Chinese characters

appropriate one for this application. It reduces code space and buffer space. It has good video scan, high speed, and highly readable low quality printing. Unfortunately this method lacks flexibility. However for all the other aforementioned reasons, the bit map method is commonly used for Korean and Chinese characters.

Because of the complexity of Korean and Chinese characters, at least a 16 by 16 resolution is required for Korean characters, and a 24 by 24 resolution for Chinese characters. 32 by 32, 64 by 64, 80 by 80, 96 by 96, and 128 by 128 resolutions are desirable when much more beauty is required and also when larger character sizes are to be produced. However, if these characters are displayed on a CRT, with 32 by 32 resolution, with the size of each character 7-10 mm square, this should be sufficient.

It is the authors' opinion that the less expensive 32 by 32 resolution CRT should be used for softcopy. The reason for this is that the price of the memory component required to hold the character definitions is continually getting less expensive. However stronger motivation is that high-speed and flexibility of typing is then possible [Ref. 1: p. 828]. IBM corporation uses 16 by 16 resolution for Gothic and 24 by 24 for Brush type Korean character syllables [Ref. 8: p. 2]. FACOM corporation uses 30 by 30 dots for Korean and Chinese characters and 24 by 30 dots for Roman alphabet, symbols, and Korean alphabet (letters) on a laser printer [Ref. 7: p. 47]. As cheaper dot matrix and laser printers find their way into the marketplace, the quality of characters will become less of an issue. Presently there are few problems with the quality of representing Korean characters that cannot be solved through the additional expenditure of money. For the definition of each character, the authors have presented two alternatives; software and hardware (character generator). In order to increase speed and usability, a hardware-oriented character generator is best. If cost and flexibility are the criterion, software-oriented character definition programs are better.

7. Memory Space for Character Definition

To represent Korean and Chinese characters, one needs to code 5,000 characters for a character definition: (2,400 Korean characters; 1,800 Chinese characters; 800 user definable characters, Roman alphabet, or symbols). If one uses a 24 bit by 24 bit matrix font size for each character, at least 360 K bytes are required (3 byte * 24 * 5,000) for character definition and 128 K bytes are required (64 K for 16 bit address * 2 byte for an address of character definition memory) for a look-up table. The total memory requirement is 488 K bytes.

A large memory space is required for the definition of the characters. Data compression of these characters can be considered for two different purposes: data transmission, and computer storage and output. Here one is mainly interested in the latter case, where the main point is the total data amount to be stored. The method of data compression of Chinese characters can be classified by using the methodology listed in Table XV [Ref. 1: p. 820].

There is a problem associated with the enlargement and alignment of character patterns. The clarity of a character depends on the size of the reproduction. If a large size is required the resolution must be high. Otherwise, stepwise zigzags appear which to some people are unbearable. Therefore, all the patterns of different font sizes must be stored. This is uneconomical. Reproducing different character sizes from the same data is desired. However, the enlargement and shrinking of character patterns from a single set of data is quite difficult, because, if the addition or the deletion of a bit by the interpolation is not done properly, it has a negative influence on the asthetics. In enlargement, the smoothness of an edge is particularly important, while in shrinking the gap between strokes must be carefully maintained.

TABLE XV

Varieties of Data Compression Methods

transmission	page unit	run-length coding
	character unit	two-dimensional predictive coding
		coding by scan line pattern unit
memory & reconstruction	dot pattern representation	coding by m by n block pattern unit
	stroke representation	checker board sampling
	contour coordinate coding	hexagonal board sampling
	contour following coding	
	mathematical equation for strokes	
enlarge/shrink	synthesis from partial character patterns	

A comparative review of the options contained in Table XV with regard to determining memory size is very difficult. This is because the requirements for character print qualities are quite different depending on each method. Simplicity in the hardware and software implementation of the compression and reconstruction of characters is a very important consideration. Generally speaking high data compression methods need complex hardware and longer times for reconstruction. Therefore, the tradeoff to be considered is between the data compression ratio and the memory size. This represents the classic economic tradeoff between the hardware/software cost with regard to the speed of character regeneration.

Because the price of the memory component is becoming less expensive the high-speed simple reconstruction method is preferred despite the necessarily large size memory. Many commercial machines have adopted this concept, and store the character dot patterns as they are without any data compression. For example, IBM machines use only a 12 by 24 font size for simple letters (Roman and Korean alphabet and symbols) instead of a 24 by 24 font [Ref. 8: p. 17]. The FACOM machines use the software definition of the second level of Korean and Chinese characters which are not used frequently for data compression [Ref. 7: p. 12]. Because of the reduction in the price of memory, the marketplace has shifted towards providing direct character storage, i.e. a large memory, instead of utilizing data compression.

V. EVALUATION OF SUGGESTED METHODS

The principal problems in current editing technology for Chinese and Korean characters were detailed in Chapter III. Fundamentally, the problems cause user inconvenience, require lengthy input procedures, and result in complex update requirements. The authors' suggested methods will solve most problems which are encountered in Korean language data processing. More research and development remain in the following areas:

First, in an enumeration method, there is no problem except low readability to Koreans and the inability to represent Chinese characters. In this case, Chinese characters are ignored because Korean language data can be represented through the use of only Korean characters without serious problems. Low readability is caused by unfamiliarity and the unbalanced shape of each letter when written by an enumeration method. With a minor change of shape of the letters, this method will eliminate the above problems.

Second, the 16-bit code for the three kinds of characters requires the consideration of the following problems:

1. The 32 key tops are complex since each key top represents three or four letters and acronyms. One solution to this problem is to use lighted, changeable key tops which represent only one letter or acronym at a moment according to the function keys and the order of keystrokes(i.e., 1st, 2nd, 3rd letter and acronym).
2. The user must remember whether a letter to be typed is the first, second, or third letter, and whether it is an acronym of a sound or a meaning character.

3. If a user does not know the meaning of a certain Chinese character to be typed, one must look up a table which shows all meanings and sounds of all possible Chinese characters.
4. In typing Korean characters which consist of only first and second letters, a user has to hit the default key to make a 16-bit code. Instead of second letter and default keys, one can use twenty one more second letter keys which generate 10-bit code as the second and third letters. Unfortunately, this will make the keyboard more complex.
5. Regardless of the authors' analysis of sound and meaning characters and careful rearrangement of these codes, duplicate codes still exist. This is because of the irregularities caused by a natural evolution of sound and meaning characters for over 2,000 years. Generally 3,000 or more Chinese characters will cause duplicate codes to increase proportionally. In order to eliminate the duplicate codes, the Korean language committee needs to take measures to clarify the meanings of the Chinese characters that cause duplicate codes to exist.

Before the actual construction of the suggested system, an economic (Cost/Benefit) analysis needs to be considered. Given the r % discount rate and the various yearly costs and benefits estimated by past data, Table XVI [Ref. 14] shows the following formula which can be used to derive the net present value of this project: This simply states that the net present value (NPV) is equal to the sum of the differences between benefits (B) and costs (C) in each year (i) of the project life (T), divided by the relevant factor (r) for that year. The current estimate of the market size for word processing in Korea is $ 2.5 million annually (Korean Daily Times, Sep 10 1984). But this estimate will be in inverse

TABLE XVI

Net Present Value Formula

$$NPV = \sum_{i=1}^{T} \frac{(B_i - C_i)}{(r)^i}$$

Legend: NPV: Net Present Value
B: Benefit
C: Cost
i: Each year
T: Project life
r: Relavent factor

proportion to the price of the system and will be in direct proportion to the usefulness and the user-friendliness until maturation.

In the above formula, the market price of the system influences the benefits for manufacturers and costs for users. This system is feasible when the net present values are positive for both manufacturers and users. If the benefits for manufacturers and the costs for users are constant in a system, the main problems will be:

1. How to minimize the costs for manufacturers
2. How to maximize the benefits for users

To solve the above problems, the best approach will be to make an efficient and user friendly system for Korean language data processing. This will increase the number (N) of systems sold, and increase the individual productivity of the users.

There are many factors and constraints which cause high cost in implementing this method. Among these, the following

three factors affect the cost performance ratio for both manufacturers and users:

1. The initial design cost: For this system, an organization has to invest initially for the design of about 5,000 Korean and Chinese character patterns, and the system software and hardware. As the number of systems produced by a manufacturer is increased, the unit cost of each system will be decreased as the costs are spread over more units.
2. Cost for character generator: As mentioned in Chapter IV, one needs about 500 K bytes memory capacity for these character definitions. The cost of memory is decreasing and the speed is increasing as technology is being develcped. This cost is an initial cost to users when buying a system.
3. Cost for hardcopy: One can consider three kinds of printer for hardcopy: dot matrix printer, chain printer, and laser printer. It is not practical to use a chain printer for our system since the chain will be approximately twenty meters long (5,000 character * 4 mm per each character) and it would be prohibitively slow. Currently, dot matrix printers and laser printers cost more than chain printers, but they are the only viable option.

Among the three kinds of cost, the third one is the most serious since the cost of hardcopy is increasing as its use increases. Recently laser printers have become more popular for these characters because of the good quality, high speed and decreasing price. Comparatively though, laser printers are still relatively expensive.

VI. RECOMMENDATION AND CONCLUSION

As the demand for data processing in Korea increases, users will continue to encounter more and more problems in utilizing the Korean language for data processing. The current methods of convergence, display and select to implement the Korean language in data processing must only be considered as interim measures due to their inefficient and time consuming means of data entry. In order to prevent this problem from becoming more complicated due to the development of various new implementations forwarded by independent research, a standardized system must be developed.

This study examined two possible solutions for using the Korean language in data processing. The enumeration method is technologically feasible, inexpensive, easy to implement, but could not be used for applications within the Korean data processing environment. This is because it results in a textual form of Hangul that is unfamiliar to most Korean people. Therefore the current enumeration method is not a feasible solution to the Korean data processing problem.

The second method examined was based on a 16-bit code representation of Korean, Chinese characters, and the Roman alphabet. This method was found to possess all the advantages currently realized by the EBCDIC or ASCII code representation of western countries. The only drawback to this system is that it might not be cost effective based on current technology. However, due to the rapid development of hardware and software technology, a cost effective means should be available within the next few years.

In order to accelerate the determination of a thorough broad based solution to the Korean data processing problem, the Korean government must organize and charge a national

level committee with the responsibility for investigating the problem and determining a viable solution. This study with its proposal of a 16-bit character code should be provided to that committee for further examination. This proposal represents a concept that could eventually lead to a long term viable solution to the data entry and processing problems of using the Korean language.

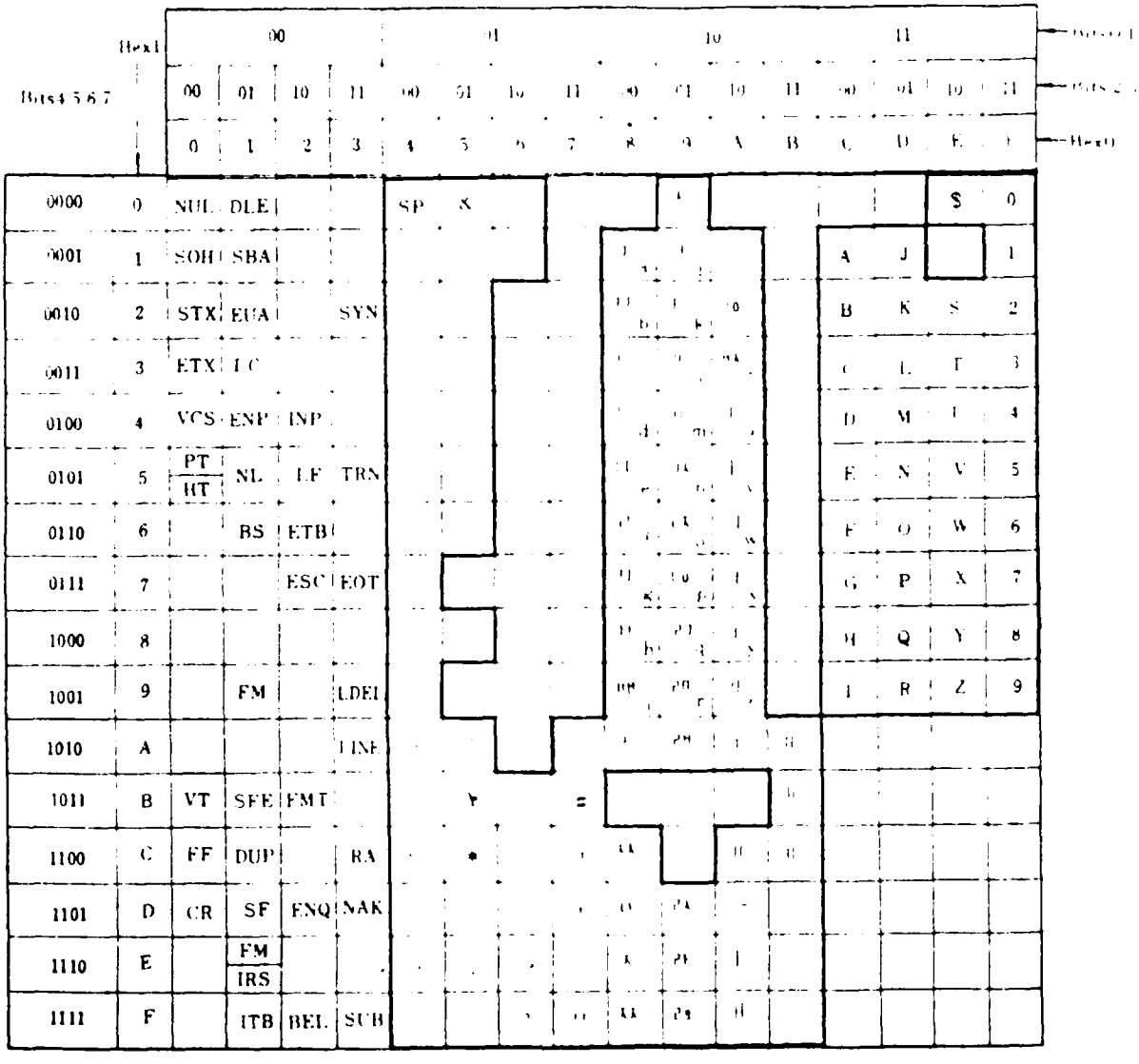
APPENDIX A

THE EVOLUTION OF CHINESE CHARACTEES

THE EVOLUTION OF CHINESE CHARACTERS

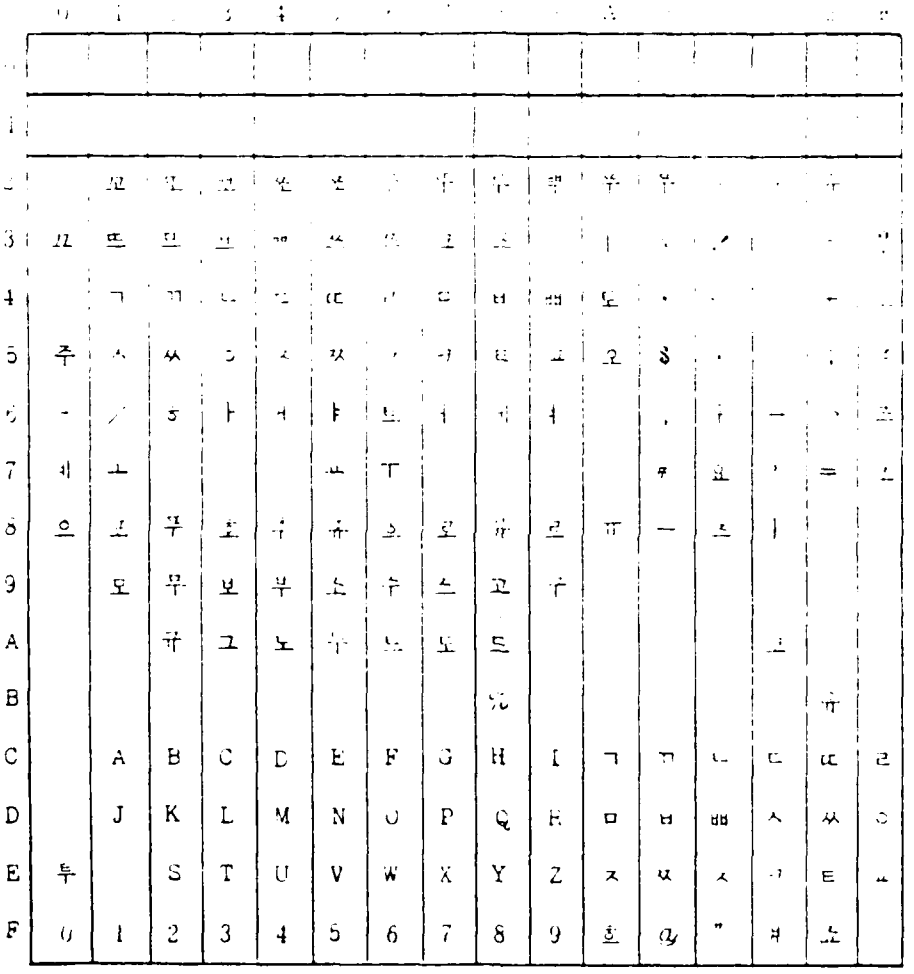
APPENDIX B

EBCDIC INPUT CODE



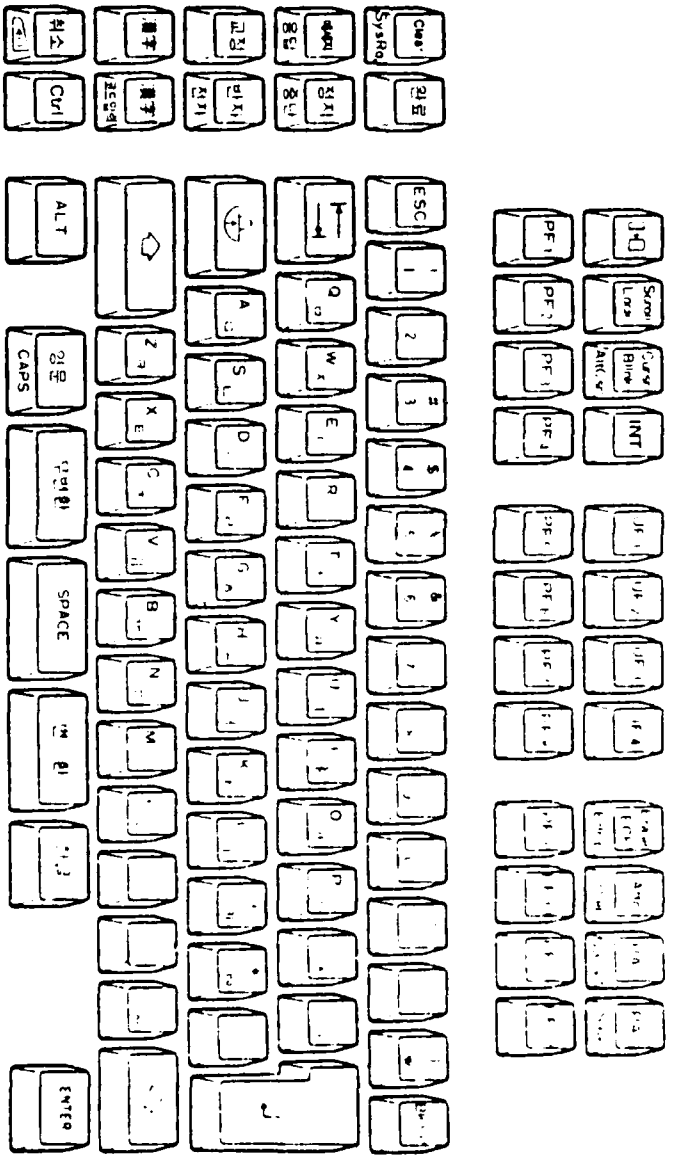
APPENDIX C

MDS INPUT CODE



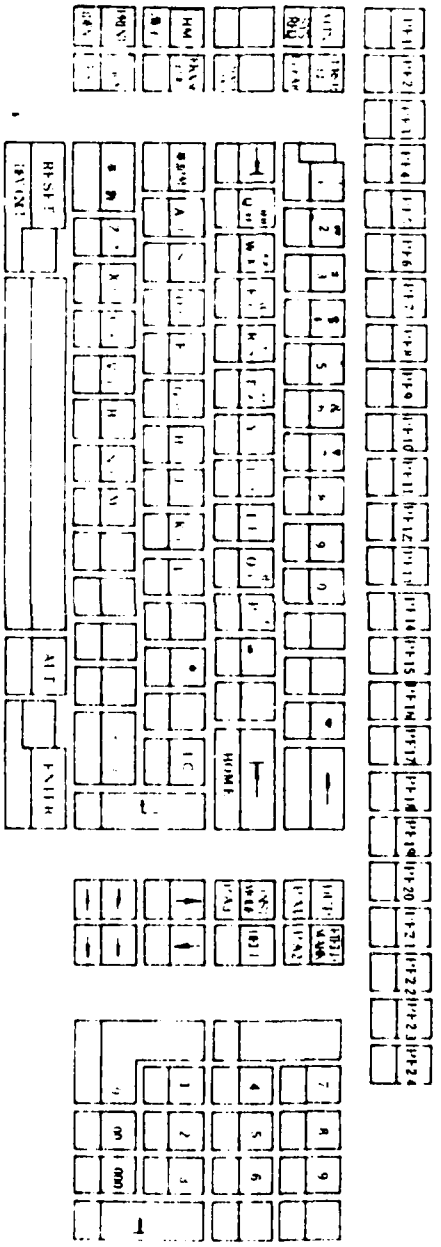
APPENDIX D

IBM MULTISTATION 5550 KEYBOARD



APPENDIX E

FACCM OS IV(KEF) KEYBOARD



APPENDIX F

LOAD COMMAND PROGRAM FOR CURRENT KEYBOARD

```
1(a"KC")
1" "00;
1"A"00081422417F4141;
1"D"007C22212121227C;
1"E"007F40407C40407F;
1"F"007F40407C404040;
1"G"001E21404047211E;
1"H"0041414174141441;
1"I"003E08080808083E;
1"J"003E41414141413E;
1"P"007E41417E404040;
1"Q"003E41414145423D;
1"R"007E41417E444241;
1"S"003E41403E01413E;
1"T"007F080808080808;
1"U"0041414141414 3E;
1"W"0041414149495522;
1"Y"0041221408080808;
1"a"00007F4141414 17F;
1"d"00001C2241412210;
1"e"00007F4040404 07F;
1"f"00007F01017F40 7F;
1"g"0000087F003E413E;
1"h"0000808080808 87F;
1"i"0000203E20203E20;
1"o"0C0042424427E242;
1"p"0000090909790909;
1"q"000041417F41417F;
1"r"00007F0101010102;
1"s"0000404040404 07F;
1"t"0000080808142241;
1"u"0000011FC1011F01;
1"w"00007F0808142241;
1"y"0000222222222 27F;
1"!"0008080808080008;
1"1"0008182808080 83E;
1"<"0001061860180601;
1"E"007E21213E21217E;
1"C"001E21404040211E;
1"J"000E0404040444 38;
1"K"00434C5860584C43;
1"L"0040404040404 07F;
1"M"0041635549414141;
1"N"0041615149454341;
1"V"0041412222141C08;
1"1"0000080808080808;
1"m"0000000000000 07F;
1"n"00007F0808080808;
1"v"00007F222222227F;
1"x"00007F007F40407F;
1"z"00007F011F010204;
1","000000000302040;
1"."000000000001818;
1">"0020180601061820;
1"'"0000221475 1422;
1"~"0000494977494977;
1":"000006087F0808;
1";"0000097909097909;
1"["00007F2222225549;
1"]"0000774444444477;
1"("0000602010;
1")"000E10107010100E;
1"#"0024247E247E2424;
1"$"00083E483E093E08;
1"%"0061520408102343;
1"&"00060408;
1"("0004020101010204;
1")"0008000000000 0FF;
1"*"001C2040404020 1C;
1"+"0018180000 1818;
1"-"0000007F;
1"/"0002222364949 49;
1"0"001C22454951221C;
1"2"003C42010E30407F;
1"3"007F02040E01413E;
1"4"00040C14247F0404;
1"5"007F407E4101413E;
1"6"001E21407E51211E;
1"7"007F0102040810 20;
1"8"003E41413E41413E;
1"9"003C42433D01423C;
1"="000C0C0000C0C0810;
1"?"003E41060800 0008;
1"@"00242424;
1"\"0000771111111122;
1"^"0038444436454639;
1"_"0000007F007F;
1"`"004141FF41495522;
```

APPENDIX G

LOAD COMMAND PROGRAM FOR CAPITAL LETTER KEYBOARD

```
1(a"lee")
1" "00;
1"!"0000007F01010101020C;
1"#"000000007F0814224141;
1"$"00000083E483E093E08;
1"%"0000000007F0022227F;
1"&"0000000000001818;
1"'"00000039090979090909;
1"("00000050525277D0505;
1")"00000090979097907909;
1"0"00000057D1525504545;
1"1"00FF01010101010101;
1"2"007F4141414141417F;
1"3"007F0808142241414141;
1"4"007F0024242424247F;
1"5"00494949497F4949497F;
1"6"0000004040407E40404C;
1"7"000C0000000808080807F;
1"8"00000000000C000007F;
1"9"00000010101010305791;
1":"00000017D1121274141;
1";"00000007C7F4141417F;
1"A"000000007F4040404075;
1"B"00000751515774577;
1"C"000000007F007F40407F;
1"D"00000007F013C010200C;
1"E"0000001C007F08142241;
1"F"00000770909091122441;
1"G"00000771515754545777;
1"H"00000007711111714172;
1"I"000000004040404007F;
1"P"0000001C007F003E413E;
1"S"00000000408142241411;
1"T"00000001212324C4949;
1"V"00000007212122549491;
1"W"0C000000417F41417F;
1"X"0000000C1C2241412210;
1"Y"00000007710157540777;
1"Z"000000007F01017F407F;
1"^"0C0000064C4F4C4649761;
1"a"007F4040404040407F;
1"c"007F00007F4040407F;
1"d"007F0101013F0101C101;
1"e"001C007F0814224141411;
1"k"00000004444447C4444;
1"l"000000101017F09091151;
1"m"000000007F1414141414;
1"n"00000002023E023E0202;
1"o"00000001111111F0101;
1"p"0000000404242770404;
1"q"00404040404040407F;
1"r"0C1C007F001C2241221C;
1"s"0C01020408142241414141;
1"t"007F2424243444494949;
1"u"000000000C141414147F;
1"v"00774444444444444477;
1"w"00414141417F4141417F;
1"x"001C2241414141221C;
1"y"00000040407C407C404C;
1"z"007F0101017F4040407F;
1"{"0000001030501010107C;
1"("0000007F02040E01413E;
1")"0000000040C14247F0404;
1"*"0000003C42010E30407E;
1"I"0000001C22405E51211E;
1"U"000007F405E51014135;
1"V"000022147F1422;
1"+"00181800001818;
1"-"0000007F;
1":"000008067F0808;
1"<"0001061860180601;
1"="000C0C00000C0810;
1">"0020180601061820;
1"?"003E41063060300C;
1"J"003E41413E41413E;
1"K"003C42433D01423C;
1"L"00384444364454639;
1"M"001C22454951221C;
1"N"00380808080800C08;
1"O"007F0102040810207;
1"P"0061520408102343;
1"Q"001020404040C2010;
1"R"00960406;
1"S"00040201010102041;
1"_"0000007F007F;
1"`"00414145F4145522;
```

APPENDIX H

IBM 2-BYTE INTERNAL HANGUL CODE

Bit-0_34	A	B	C
00000	Not Used-1	Restricted-1	Restricted-1
00001	Full Code-1	Restricted-1	Full Code-1
00010	ㄱ	Full Code-1	
00011	ㄲ	ㄱ	
00100	ㄴ	ㄲ	ㄱ
00101	ㄷ	ㄴ	
00110	ㄸ	ㄷ	ㄴ
00111	ㄹ	ㄸ	ㄷ
01000	ㅁ	Restricted-1	ㄸ
01001	ㅂ	Restricted-1	ㄹ
01010	ㅃ	ㅁ	ㄺ
01011	ㅅ	ㅂ	ㄻ
01100	ㅆ	ㅃ	ㄼ
01101	ㅇ	ㅅ	ㄽ
01110	ㅈ	ㅆ	ㄾ
01111	ㅉ	ㅇ	ㄿ
10000	ㅊ	Restricted-1	ㅀ
10001	ㅋ	Restricted-1	ㅁ
10010	ㅌ	ㅈ	ㅂ
10011	ㅍ	ㅉ	ㅄ
10100	ㅎ	ㅊ	ㅅ
10101	Not Used-3	ㅋ	ㅆ
10110	Not Used-3	ㅌ	ㅇ
10111	Not Used-3	ㅍ	ㅈ
11000	Not Used-3	Restricted-1	ㅊ
11001	Not Used-3	Restricted-1	ㅋ
11010	Not Used-3	ㅎ	ㅌ
11011	Not Used-3	ㅏ	ㅍ
11100	Not Used-3	ㅐ	ㅎ
11101	Not Used-3	ㅑ	ㅏ
11110	Not Used-2	Not Used-3	Not Used-3
11111	Restricted-1	Not Used-3	Restricted-1

LIST OF REFERENCES

1. Nagao, Makoto, "Data Compression of Chinese Character Patterns", Proceedings of the IEEE, Vol. 68, No. 7, PP. 816-819, July 1980.

2. Mackenzie, C. E., Coded Character Sets, History and Development, Addison-Wesley Publishing Company, 1980.

3. Dong-A Publishing Company, 1800 Characters of Sino-Korean Textbook, 1978.

4. Department of Computer Engineering., Seoul National University, English/Hanguel/Chinese Character Display Cortroller Design Using Address Conversion Technique and DMA, by C. M. Kim and H. Y. Hwang, 17 May 1982.

5. Rev. Kim, Jacob Chang-Ui, "The Study of Etymology, Classification and Signification of the Chinese Character", Sino-Korean, Herald Printers, Monterey, California 1984.

6. Lee, M. N., A Study on 2-Byte Hangeul Database Systems, M.S. Thesis, Yonsei University, Seoul, June 1982.

7. FACOM OS IV KEF Manual, 1st ed, V.1, FACOM Korea Corp., June 1982.

8. IBM Multistation Manual, 1st ed, IBM Korea Corp., 1984.

9. Grant, Bruce K., A Guide to Korean Characters, Hollym International Corp., 1979.

10. Defense Language Institute Foreign Language Center, Korean: Resource Module I Sounds and Hangul, September 1980.

11. Hangul Academy, A History of Hangul Academy, December 1981.

12. Coueignoux, "Character Generation by Computer", Computer Graphics and Image Processing, pp. 240-269, 1981.

13. Foley, James D. and Van Dam, Andries, Fundamentals of Interactive Computer Graphics, Addison-Wesley Publishing Company, March 1983.

14. Fin-Dor, Phillip and Jones, Carl R., Information Resources Management (Draft).

BIBLIOGRAPHY

Hamacher, V. Carl, Vranesic, Zvonko G., and Zaky, Safwat G., Computer Organization, McGraw-Hill Book Company, 1978.

Hall, Douglas V., Microprocessors and Digital Systems, 2nd ed., McGraw-Hill Book Company, 1983.

Madnick, Stuart E. and Donovan, John J., Operating Systems, McGraw-Hill Book Company, 1974.

Panco, Raymond R., "Kanji Keyboard Chaos", Computerworld, 28 November 1983.

People's Republic of China Library Academy, Chinese Character Code for Information Interchange, 1969.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Technical Information Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, Code 0142 Naval Postgraduate School Monterey, Califcrnia 93943	2
3.	Department Chairman, Code 54 Department of Administrative Sciences Naval Postgraduate School Monterey, Califcrnia 93943	1
4.	Department Chairman, Code 52 Hq Department of Ccmputer Science Naval Postgraduate School Monterey, California 93943	1
5.	Professor Michael J. Zyda, code 52 Department of Ccmputer Science Naval Postgraduate School Monterey, Califcrnia 93943	1
6.	Professor Carl R. Jones, code 54 Js Department of Administrative Sciences Naval Postgraduate School Monterey, Califcrnia 93943	1
7.	Library Officer Korea Military Academy Seoul, Korea 130-09	2
8.	Office of the Defense Attache Embassy of the Republic of Korea 2320 Massachusetts Avenue, Northwest Washington D.C. 20008	1
9.	Major Chong Hae Kim Planning and Management Staff Department Republic of Korean Army Headquarters Seoul, Korea 130-09	5
10.	Major Sung Woo Kc Planning and Management Staff Department Republic of Korean Army Headquarters Seoul, Korea 130-09	5
11.	Major Jang Hun Lee, SMC 1366 Naval Postgraduate School Monterey, California 93943	1
12.	Captain Hee Young Lee, SMC 2726 Naval Postgraduate School Monterey, Califcrnia 93943	1
13.	Captain Hwa Dal Song, SMC 2748 Naval Postgraduate School Monterey, Califcrnia 93943	1

14. Captain Kwang Jun Choi, SMC 2870 — 1
Naval Postgraduate School
Monterey, Califcrnia 93943

15. Captain Jai Eur Jang, SMC 1032 — 1
Naval Postgraduate School
Monterey, Califcrnia 93943

16. Joon Yong Lee — 1
Department of Korean Language
Defense Language Institute
Presidio of Monterey, California 93940

17. Academic Library — 1
Defense Language Institute
Presidio of Monterey, California 93940

18. Professor Young S. Shin, code 69 Sg — 1
Department of Mechanical Engineering
Naval Postgraduate School
Monterey, California 93943

END

FILMED

5-85

DTIC